Supplementary Information

Genetic and transcriptional evolution alters cancer cell line drug response

Uri Ben-David[1], Benjamin Siranosian[1], Gavin Ha[1,2], Helen Tang[1], Yaara Oren[1,3], Kunihiko Hinohara[1,2], Craig A. Strathdee[1], Joshua Dempster[1], Nicholas J. Lyons[1], Robert Burns[2], Anwesha Nag[2], Guillaume Kugener[1], Beth Cimini[1], Peter Tsvetkov[1], Yosef E. Maruvka[1], Ryan O'Rourke[1,2], Anthony Garrity[1], Andrew A. Tubelli[1], Pratiti Bandopadhayay[1,2,3], Aviad Tsherniak[1], Francisca Vazquez[1], Bang Wong[1], Chet Birger[1], Mahmoud Ghandi[1], Aaron R. Thorner[2], Joshua A. Bittker[1], Matthew Meyerson[1,2,3], Gad Getz[1,5], Rameen Beroukhim[1,2,3,4,#] and Todd R. Golub[1,2,3,6,#]

[1] Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA

[2] Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[3] Harvard Medical School, Boston, Massachusetts, USA

[4] Brigham and Women's Hospital, Boston, Massachusetts, USA

[5] Massachusetts General Hospital, Boston, Massachusetts, USA

[6] Howard Hughes Medical Institute, Chevy Chase, Maryland, USA

[#] These authors jointly supervised this work.

Correspondence should be addressed to R.B. (rameen_beroukhim@dfci.harvard.edu) or T.R.G. (golub@broadinstitute.org)

Table of contents

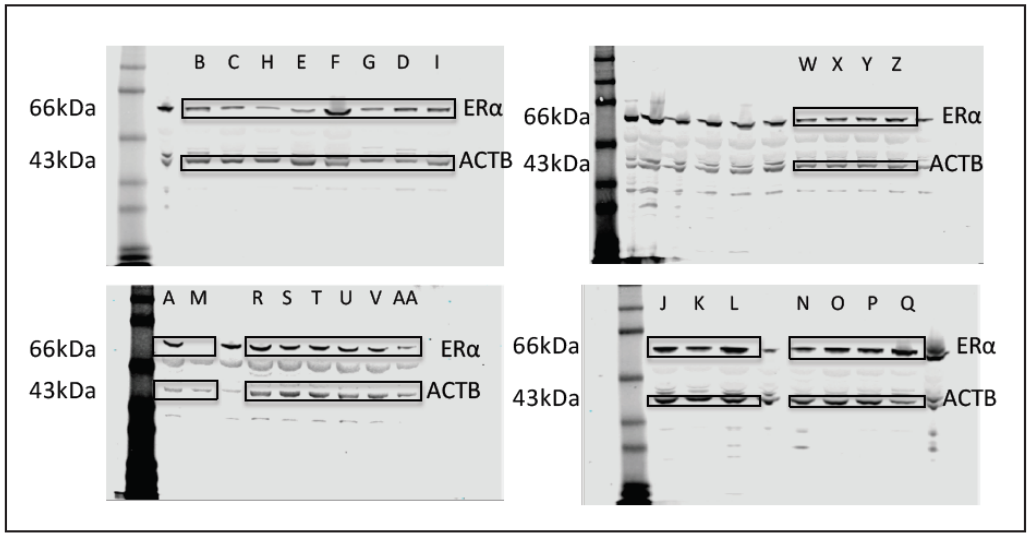
Supplementary Figure:

Supplementary Discussion:

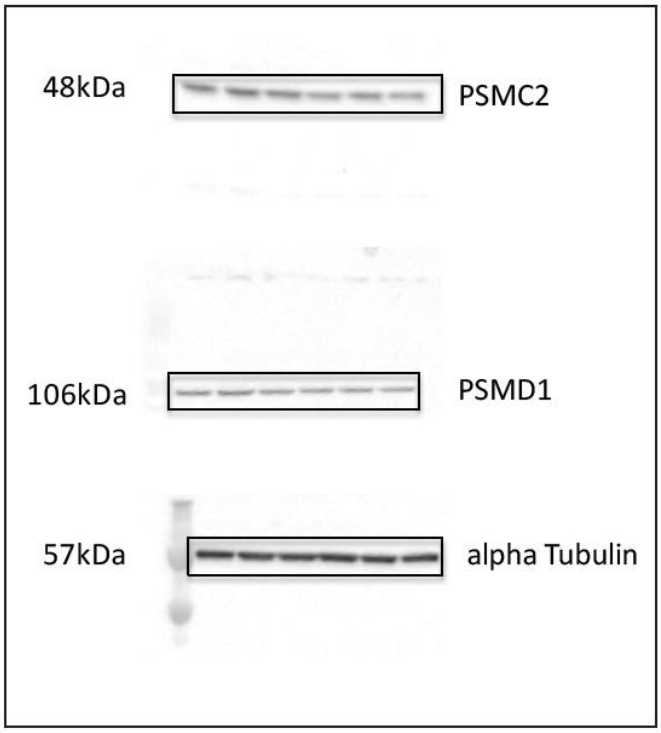
Supplementary References

Supplementary Figure 1

Extended Data Fig. 3b



Extended Data Fig. 10n



Supplementary Discussion

1. MCF7 is a common, representative cancer cell line

To more systematically address whether the observed subclonality of cell lines is stable, or can result in clonal selection, we focused on MCF7, a cell line that has been used as a model for estrogen receptor (ER)-positive breast cancer in over 30,000 publications[12-14], and whose genomic characteristics are representative of cell lines within the CCLE[4] (**Extended Data Fig. 1**). While examples of genetic or phenotypic drift among MCF7 and other cultures have been reported[62-77], analyses have been limited, and the impact of such genetic variation on drug response has not been explored.

2. ESR1 copy number loss is associated with reduced expression of its protein product

A strong correlation was observed between the copy number status of ESR1, and ERα protein expression. Interestingly, despite a normal copy number of *ESR1*, Strain M, which was selected for tamoxifen resistance, had the lowest protein levels of the *ESR1* protein product across the MCF7 panel (**Extended Data Fig. 3b-c**), indicating that *ESR1* loss occurred via a different mechanism.

3. Variation in chromosomal translocations across MCF7 strains

Chromosomal translocations were highly variable across strains: a median of 12 gene rearrangements (range 7 to 17) were identified per strain, with the majority (55%) detected in only a subset of strains, and 40% detected in only a single strain.

4. Genomic evolution of cell lines appears to be different from genomic evolution of tumors during disease progression

The genetic changes that arise during the propagation of cell lines in culture may be associated with actual breast cancer disease progression. To address this question, we compared the genetic changes that arose following MCF7 passaging to those that arise during breast cancer disease progression. Yates et al.[78] found seven genes that were significantly more frequently mutated in relapsed or metastatic breast cancer than in primary breast cancer. In a pairwise comparison of earlier vs. later passage MCF7 strains, we could not detect new mutations in any of these genes.

We recently reported that recurrent arm-level CNAs are seen more frequently in relapsed or metastatic tumors compared to matched primary tumors across cancer types, but that some of these events tend to gradually disappear during the propagation of patient-derived xenografts[53]. In breast cancer, the number of chromosomal changes was reported to be higher in metastases than in primary tumors, and recurrent arm-level CNAs (e.g., loss of 18q) were reported to be more frequent in metastases than in their matched primary tumors[79-82]. In contrast, the number of arm-level CNAs did not increase following extensive passaging of MCF7 ($p=0.31$), and – similar

to our previous findings in PDXs – common CNAs mostly disappeared rather than arose on prolonged cell line passaging (p=0.047 in a McNemar's test).

Therefore, we found no evidence that genomic evolution in culture reflected breast cancer disease progression in patients. We note that tumors and cell lines probably follow distinct evolutionary trajectories due to the widely distinct *in vivo* and *in vitro* selection pressures to which they are subjected.

5. Single cell RNA sequencing following MCF7 response to bortezomib

To validate that scRNA-seq could be used to evaluate transcriptional changes in MCF7-derived populations, we treated MCF7 (strain AA) with bortezomib (500nM), and profiled the transcriptome of 7,555 single cells at four time points: before drug treatment (t0), 12 hours after drug exposure (t12), 48 hour after drug exposure (t48), and 72 hours after drug exposure followed by 24 hours without it (t72+24). Cell populations clustered separately at the different time points (**Extended Data Fig. 6j**). In line with the previously reported transcriptional response to bortezomib[83-85], drug treatment led to upregulation of the proteasome and unfolded protein response gene expression programs, and to downregulation of estrogen signaling and proliferation signatures (**Extended Data Fig. 6k-n**). As expected, and in line with what we observed visually (data not shown), the genes most differentially expressed at t48 vs. t0 were significantly enriched for apoptosis ($p<2*10^{-16}$). We conclude that scRNA-seq can successfully capture transcriptional differences between MCF7 cell populations.

6. Single cell RNA sequencing identifies differentially expressed genes between single cell-derived clones

Despite an overall similarity between the single cell clones and their parental populations, a median of 165 differentially expressed genes (DEGs) (|log2FC|>0.25, Bonferroni-corrected p<0.01; range, 95 to 226) were identified in a pairwise comparison (**Supplementary Table 15**). Almost twice as many genes were identified between the clonal populations themselves (median: 311; range, 290 to 395), indicating that single cell clones were closer to their parental population than to each other. Moreover, 153 DEGs were detected even between the two cultures of the same clone, before and after prolonged passaging (**Supplementary Table 15**). Moreover, cultures of the same clone before and after prolonged passaging exhibited significant differences in important pathways, including estrogen signaling and proliferation (**Extended Data Fig. 6r** and **Supplementary Table 15**).

7. Genetic variation across strains of additional cancer and non-cancer cell lines

MCF7 has been in culture for decades, and may have been subjected to varying selection pressures. We therefore tested the genetic variation of multiple additional cell lines. In paired comparisons between strains of the same cancer cell line, a median of 17% of the detected non-silent mutations (range, 0% to 33%) were only identified in one strain (**Extended Data Fig. 8a**,**g**

and **Supplementary Table 23**). The variation within our MCF7 and A549 panels was well within this observed range, with MCF7 being more variable than the median cell line (median discordance, 27%; range, 0% to 44%) and A549 being as variable as the median cell line (median discordance, 15%; range, 2% to 30%) (**Extended Data Fig. 8a**). The variation between the two samples of the benign tumor cell line BEN-MEN-1 – separated by three years of continuous culture – was not lower than that within malignant cell lines (**Extended Data Fig. 8a**,**g**), indicating that genomic instability is not limited to cell lines from malignant tumors.

We analyzed multiple strains of the non-transformed human cell lines MCF10A[25], HA1E[86] and RPE1 (ref. [87]). Whereas the variation in non-silent mutations across kidney HA1E strains was relatively low (median discordance, 5%; range, 2% to 7%), the variation across breast MCF10A strains was as high as that seen in MCF7 cancer cells (median discordance, 26%; range, 17% to 40%; **Extended Data Fig. 8a**,**h** and **Supplementary Table 23**). Similarly, re-analysis of genomic data from immortalized RPE1 retinal cells showed discordant arm-level CNAs in three of 26 strains (~10%) (**Extended Data Fig. 8b** and **Supplementary Table 24**).

To ask whether the degree of cell line genomic instability is associated with its degree of transformation, we compared MCF10A cells at different stages of transformation[88-90]. We compared fraction of discordant non-silent SNVs between non-transformed MCF10A strains (expressing GFP or an empty vector), partially-transformed MCF10A strains (expressing a single oncogene), and a fully-transformed MCF10A strain (expressing ERBB2 and 1433ζ; **Supplementary Table 21**). The fully-transformed MCF10A strain was significantly more unstable than its partially-transformed or non-transformed counterparts (**Extended Data Fig. 8c**,**h**).

Taken together, these results suggest that genomic instability is not restricted to transformed cells, but that increasing instability is associated with malignant progression. A causal relationship, however, cannot be inferred from these analyses.

8. Doubling time is strongly correlated to the number of protein altering mutations and the extent of subclonality of the strains

Because mutation load serves as a molecular reflection of cellular "age"[91], we examined the association between mutation load and growth rate. Remarkably, doubling time was strongly correlated with the number of protein altering mutations in naturally-occurring strains (Spearman's Rho=-0.79, $p=1*10^{-4}$; **Extended Data Fig. 9h**), consistent with cell line fitness increasing with time in culture. Similarly, the extent of subclonality of strains was highly predictive of proliferation rate (Spearman's Rho=-0.77, $p=2*10^{-4}$; **Extended Data Fig. 9i**), consistent with the association between subclonality and prognosis observed in primary tumors[92].

9. Cell line instability is associated with the degree of aggressiveness and genomic instability of its tumor of origin

Analysis across 106 cell lines common to the Broad CCLE and Sanger GDSC indicated that the concordance in mutation allelic fractions was significantly lower for cell lines derived from metastases as opposed to primary tumors (**Extended Data Fig. 8d**). Indeed, cell lines derived from metastases exhibited significantly higher genomic instability scores[57,93] than those derived from primary tumors (**Extended Data Fig. 8e**). Of note, the concordance between mutation AFs was also significantly lower for cell lines with microsatellite instability (MSI), suggesting that deficient maintenance of genome integrity underlies at least some of the observed differences (**Extended Data Fig. 8f**). These data suggest that cell line instability is associated with the degree of aggressiveness and genomic instability of its tumor of origin.

10. Variability in drug response is observed in multiple analyses

The high variability in drug response across MCF7 strains was observed when the analysis was performed in multiple ways: 1) When broadly resistant strains were removed (**Extended Data Fig. 10c**); 2) when using a more stringent definition of strong activity (>80% growth inhibition; **Extended Data Fig. 10d**), and 3) when averaging the 10 WT Connectivity Map strains (**Extended Data Fig. 10j**).

11. The Connectivity Map strains do not skew variation estimates

As our panel of MCF7 strains included 11 CMap strains, which were highly similar to one another, we wanted to confirm that our estimations of variation across strains were not skewed due to this closely-related group of strains. We therefore averaged the genetic, transcriptomic and drug response profiles of the 10 naturally-occurring CMap strains, and repeated the quantitative analyses of variation using this single averaged CMap strain. The degree of heterogeneity across strains was very similar to that observed when the CMap strains were considered independently. This is true for the variation in the SNV and CNA landscapes (**Extended Data Fig. 10e-f**), the correlations between doubling time and the number of protein altering mutations (**Extended Data Fig. 10g**) and between doubling time and subclonal fraction (**Extended Data Fig. 10h**), the variation in gene expression (**Extended Data Fig. 10i**), and the variation in drug response (**Extended Data Fig. 10j**).

12. Dose-response curves are indicative of heterogeneous drug response

Dose-response analysis showed that while some compounds were active in all strains, their EC50 often varied dramatically. For example, the histone deacetylase inhibitor romidepsin varied in its cytotoxicity by more than 800-fold across the strains (**Extended Data Fig. 10k**). We note that the dose-response curves of many differentially active drugs were characterized by shallow hill slopes and incomplete killing in the resistant strains (**Extended Data Fig. 10k**), which is indicative of heterogeneity within the cell population[94].

13. Comparison of clustering trees

Our results suggest an overarching view in which genetic variation across strains of common cell lines is associated with relevant variation in cell morphology, gene expression, and drug response (**Fig. 1g**, **2a**, **3b, 4a** and **Extended Data Fig. 9g**). We compared the MCF7 clustering trees representing these different datasets using the Fowlkes-Mellows index[52] (see **Methods**), and found that all trees were much more similar to one another than expected by chance (**Extended Data Fig. 11a**). The similarity between the genetic trees and the gene expression tree was higher than that between either of them and the drug response tree (**Extended Data Fig. 11a**). Among the genetic trees, the arm-level copy number-based tree was the most similar to the gene expression-based tree (**Extended Data Fig. 11a**), probably due to the direct effect of aneuploidy on the expression of hundreds of genes simultaneously. Finally, the drug response tree was slightly more similar to the gene expression tree than to the mutation and gene-level copy number trees (**Extended Data Fig. 11a**), in line with previous reports showing gene expression as a strong predictor of drug dependencies[95].

14. Genomic variation results in differential genetic dependencies

We asked whether genomic variation between strains affects genetic dependencies in a similar way to its effect on drug response. This is particularly important because pooled CRISPR/Cas9 knock-out screens use genome copy number estimates to correct for the non-specific toxicity of Cas9, which is associated with the number of Cas9-mediated cuts in DNA[49,96,97]. We used the CERES algorithm to generate gene dependency scores[52] from genome-wide CRISPR screens performed in both MCF7 and the cell line KPL1, now known to be a strain of MCF7 (ref. [98]). The dependency landscapes of MCF7 and KPL1 were more similar to one another than to other breast cancer cell lines (**Extended Data Fig. 11b**), but MCF7 had nearly twice as many dependencies as KPL1 (254 vs. 142, respectively; excluding pan-essential genes required for the survival of all cell lines). ~70% of the MCF7 dependencies and ~30% of the KPL1 dependencies were not shared by both cell lines (**Extended Data Fig. 11c**). Remarkably, two key genes involved in estrogen signaling, *ESR1* and *BCAS3*, were among the top differential dependencies (**Supplementary Table 31**). Moreover, the genes that were over-expressed in MCF7 compared to KPL1 were significantly enriched for estrogen signaling expression signatures (**Extended Data Fig. 11d**), which were in turn strongly associated with the dependency of breast cancer cell lines on ESR1 (**Extended Data Fig. 11e**). Similarly, the protein levels of *GATA3*, one of the most frequently mutated genes in breast cancer[99], were higher in MCF7 than in KPL1, and MCF7 was more sensitive to *GATA3* knockout (**Extended Data Fig. 11f**). We conclude from these analyses that genetic variation within cancer cell lines can have substantial impact on the interpretation of genetic loss-of-function screens.

15. Single cell-derived clones remain phenotypically unstable

Lastly, we performed functional analyses of parental strains and their single cell-derived clone, comparing their proliferation rates and their dependency on estrogen. This analysis demonstrated that single cell-derived clones remain phenotypically unstable (**Extended Data Fig. 11g-i**).

While single cell cloning initially reduces genomic heterogeneity, it does not ultimately yield genomically and phenotypically stable clonal lines.

16. Differential drug response can be used to identify drugs' mechanisms of action

Comparison of gene expression profiles between the most sensitive and most resistant cell lines or cell strains identified differential expression of the correct associated pathways in about two thirds of the cases in the MCF7 panel, compared to about a third of the cases in the CTD^2 and the GDSC panels ($p=0.022$; **Fig. 4l** and **Supplementary Table 32**). Therefore, panels of multiple strains of the same cell line, like the one described in this study, may offer a novel approach to identifying genotype-specific cancer vulnerabilities. However, we note that the utility of this approach may be limited to those drugs for which there is a highly variable sensitivity within the panel, and may be further limited by the transient nature of cell line strains

17. Comparison of naturally-occurring and genetically-manipulated strains

We found that "neutral" genetic manipulations, which usually involve antibiotic selection, are a major contributor to genomic variation across cell line strains. To ask whether genetic manipulations, as a group, lead to any reproducible transcriptional and drug response alterations, we compared the naturally-occurring to the genetically-manipulated strains. We first compared gene expression profiles between these groups and found no evidence for consistent gene expression changes. The number of differentially expressed genes between in-group pairs of strains was not statistically different from that between out-group pairs of strains. The genes differentially expressed (>2FC) between the two groups were not enriched for any hallmark pathway. We next compared the drug response patterns between the two groups. We did not find any drug that differentially ($p<0.05$, $q<0.25$) affected manipulated strains compared to naturally-occurring strains. We note that this does not rule out the possibility that specific genetic manipulations, or specific viral integration sites, may affect the genomic stability of cancer cell lines or may be associated with reproducible genomic changes. However, many more samples will be required to assess the effect of specific genetic manipulations with sufficient statistical power.

Supplementary References

62. Jones, C. *et al.* Comparative genomic hybridization reveals extensive variation among different MCF-7 cell stocks. *Cancer Genet Cytogenet* **117**, 153-8 (2000).
63. Nugoli, M. *et al.* Genetic variability in MCF-7 sublines: evidence of rapid genomic and RNA expression profile modifications. *BMC Cancer* **3**, 13 (2003).
64. Leung, E., Kannan, N., Krissansen, G.W., Findlay, M.P. & Baguley, B.C. MCF-7 breast cancer cells selected for tamoxifen resistance acquire new phenotypes differing in DNA content, phospho-HER2 and PAX2 expression, and rapamycin sensitivity. *Cancer Biol Ther* **9**, 717-24 (2010).
65. Coser, K.R. *et al.* Antiestrogen-resistant subclones of MCF-7 human breast cancer cells are derived from a common monoclonal drug-resistant progenitor. *Proc Natl Acad Sci U S A* **106**, 14536-41 (2009).
66. Esquenet, M., Swinnen, J.V., Heyns, W. & Verhoeven, G. LNCaP prostatic adenocarcinoma cells derived from low and high passage numbers display divergent responses not only to androgens but also to retinoids. *J Steroid Biochem Mol Biol* **62**, 391-9 (1997).
67. Briske-Anderson, M.J., Finley, J.W. & Newman, S.M. The influence of culture time and passage number on the morphological and physiological development of Caco-2 cells. *Proc Soc Exp Biol Med* **214**, 248-57 (1997).
68. Yu, H., Cook, T.J. & Sinko, P.J. Evidence for diminished functional expression of intestinal transporters in Caco-2 cell monolayers at high passages. *Pharm Res* **14**, 757-62 (1997).
69. Wenger, S.L. *et al.* Comparison of established cell lines at different passages by karyotype and comparative genomic hybridization. *Biosci Rep* **24**, 631-9 (2004).
70. Sambuy, Y. *et al.* The Caco-2 cell line as a model of the intestinal barrier: influence of cell and culture-related factors on Caco-2 cell functional characteristics. *Cell Biol Toxicol* **21**, 1-26 (2005).
71. Lin, H.K. *et al.* Suppression versus induction of androgen receptor functions by the phosphatidylinositol 3-kinase/Akt pathway in prostate cancer LNCaP cells with different passage numbers. *J Biol Chem* **278**, 50902-7 (2003).
72. Riley, S.A., Warhurst, G., Crowe, P.T. & Turnberg, L.A. Active hexose transport across cultured human Caco-2 cells: characterisation and influence of culture conditions. *Biochim Biophys Acta* **1066**, 175-82 (1991).
73. Langeler, E.G., van Uffelen, C.J., Blankenstein, M.A., van Steenbrugge, G.J. & Mulder, E. Effect of culture conditions on androgen sensitivity of the human prostatic cancer cell line LNCaP. *Prostate* **23**, 213-23 (1993).
74. Engelholm, S.A. *et al.* Genetic instability of cell lines derived from a single human small cell carcinoma of the lung. *Eur J Cancer Clin Oncol* **21**, 815-24 (1985).
75. Masramon, L. *et al.* Genetic instability and divergence of clonal populations in colon cancer cells in vitro. *J Cell Sci* **119**, 1477-82 (2006).
76. Frattini, A. *et al.* High variability of genomic instability and gene expression profiling in different HeLa clones. *Sci Rep* **5**, 15377 (2015).
77. Orth, K. *et al.* Genetic instability in human ovarian cancer cell lines. *Proc Natl Acad Sci U S A* **91**, 9495-9 (1994).
78. Yates, L.R. *et al.* Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell* **32**, 169-184 e7 (2017).
79. Nishizaki, T. *et al.* Genetic alterations in lobular breast cancer by comparative genomic hybridization. *Int J Cancer* **74**, 513-7 (1997).
80. Li, H. *et al.* A preliminary study of the relationship between breast cancer metastasis and loss of heterozygosity by using exome sequencing. *Sci Rep* **4**, 5460 (2014).

81. Friedrich, K. *et al.* Chromosomal genotype in breast cancer progression: comparison of primary and secondary manifestations. *Cell Oncol* **30**, 39-50 (2008).
82. Kroigard, A.B. *et al.* Clonal expansion and linear genome evolution through breast cancer progression from pre-invasive stages to asynchronous metastasis. *Oncotarget* **6**, 5634-49 (2015).
83. Meiners, S. *et al.* Inhibition of proteasome activity induces concerted expression of proteasome genes and de novo formation of Mammalian proteasomes. *J Biol Chem* **278**, 21517-25 (2003).
84. Mitsiades, N. *et al.* Molecular sequelae of proteasome inhibition in human multiple myeloma cells. *Proc Natl Acad Sci U S A* **99**, 14374-9 (2002).
85. Powers, G.L., Ellison-Zelski, S.J., Casa, A.J., Lee, A.V. & Alarid, E.T. Proteasome inhibition represses ERalpha gene expression in ER+ cells: a new link between proteasome activity and estrogen signaling in breast cancer. *Oncogene* **29**, 1509-18 (2010).
86. Hahn, W.C. *et al.* Creation of human tumour cells with defined genetic elements. *Nature* **400**, 464-8 (1999).
87. Bodnar, A.G. *et al.* Extension of life-span by introduction of telomerase into normal human cells. *Science* **279**, 349-52 (1998).
88. Nieminen, A.I., Partanen, J.I., Hau, A. & Klefstrom, J. c-Myc primed mitochondria determine cellular sensitivity to TRAIL-induced apoptosis. *EMBO J* **26**, 1055-67 (2007).
89. Lu, J. *et al.* 14-3-3zeta Cooperates with ErbB2 to promote ductal carcinoma in situ progression to invasive breast cancer by inducing epithelial-mesenchymal transition. *Cancer Cell* **16**, 195-207 (2009).
90. Seton-Rogers, S.E. *et al.* Cooperation of the ErbB2 receptor and transforming growth factor beta in induction of migration and invasion in mammary epithelial cells. *Proc Natl Acad Sci U S A* **101**, 1257-62 (2004).
91. Bromham, L. & Penny, D. The modern molecular clock. *Nat Rev Genet* **4**, 216-24 (2003).
92. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* **22**, 105-13 (2016).
93. Burrell, R.A. *et al.* Replication stress links structural and numerical cancer chromosomal instability. *Nature* **494**, 492-496 (2013).
94. Xia, X., Owen, M.S., Lee, R.E. & Gaudet, S. Cell-to-cell variability in cell death: can systems biology help us make sense of it all? *Cell Death Dis* **5**, e1261 (2014).
95. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576 e16 (2017).
96. Aguirre, A.J. *et al.* Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov* **6**, 914-29 (2016).
97. Munoz, D.M. *et al.* CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. *Cancer Discov* **6**, 900-13 (2016).
98. Capes-Davis, A. *et al.* Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int J Cancer* **127**, 1-8 (2010).
99. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).